Large Language Models (LLMs) to Identify Toxicity in the Digital Sphere during Protest Events in Latin America

Bastián González-Bustamante
Leiden University & Universidad Diego Portales
✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation delivered at the XII Latin American Congress of Political Science (ALACIP)
Lisbon, Portugal, July 17-20, 2024

Table of Contents

Introduction

Introduction

This article benchmarked the ability of **16 open-source Large Language Models (LLMs)** to perform annotation tasks on political content.

The models were deployed locally to identify **toxicity and incivility in the digital sphere** on a novel protest event dataset comprising ∼5M digital interactions.

- It constitutes a contribution demonstrating the potential of generative artificial intelligence to automate the labelling processes of political content.

Related Work

Text-as-Data and Perspective API

The text-as-data approach appears as a relevant cornerstone due to the limited number of texts that human coders can annotate (González-Bustamante, 2023; Schmidt et al., 2024).

The **Perspective API**, developed by Jigsaw and Google, is one of the top-shelf options for classifying toxicity and incivility. It was trained on a variety of other sources labelled by crowdsource raters using distilled BERT models.

In academic research, the algorithm has been used to detect toxicity and uncivil comments on Twitter (Hopp et al., 2020; Orchard et al., 2024; Schmidt et al., 2024), Facebook (Hopp et al., 2020; Kim et al., 2021; Schmidt et al., 2024), news comments (Orchard et al., 2024; Schmidt et al., 2024) and Wikipedia content (Pavlopoulos et al., 2020).

Open-source LLMs

Researchers have been using GPTs through OpenAI's API for a variety of tasks (Gilardi et al., 2023; He et al., 2024), and despite different **concerns related to reproducibility, privacy and openness** of this pay-per-use form, this way tends to offer resources beyond those usually available to the average researcher in social sciences fields, it is straightforwardly to deploy without excessive computational requirements.

Recent studies have shown that some open-source LLMs are able to perform or even surpass GPT-3.5 and 4 in zero-shot tasks, only lagging in fine-tuned GPTs (Alizadeh et al., 2024). In addition, some open-source so-called **Small Language Models (SLMs)**, $\leq 7B$ parameters, thanks to quantisation techniques, are able to perform well and offer a considerable cost reduction (Irugalbandara et al., 2024, see also Bucher and Martini, 2024).

Task Description

Data and Classification Task

Novel dataset on protest events

Protests against the coronavirus and judicial reform (August 2020)

Protests against education budget cuts (May 2019)

Social outburst in Chile (October 2019)

∼5M messages in total

LLMs zero-shot classification

› Random, balanced (on toxicity quintiles) sample of **1,000 messages** from Argentina and Chile

Open-source LLMs deployed locally

› Minimum temperature to ensure replicability*

Intel Core i9-14900K CPU

NVIDIA GeForce RTX 4070 Super Windforce OC 12GB GPU

64 GB of RAM (DDR5-4800)

Windows Subsystem for Linux v2.1.5.0 and Ollama v0.1.44

* There is no agreement about the role of this on models' hallucinations.

Gold Standard in Progress

README | Code of conduct | CC-BY-4.0 license | GPL-3.0 license

gold-standard-toxicity

Gold Standard for Toxicity and Incivility Project
Annotated Data in Spanish for Toxicity and Insults in Digital Social Networks

coverage 68% | version v0.3.3 | repo status Active | DOI 10.5281/zenodo.12574288 | license GNU GPLv3 | license CC BY 4.0
made with Raspberry Pi 5 | supported by tDL

Overview

This repository contains data sets and materials for a gold standard elaboration on toxicity and incivility in the digital sphere based on human coding to benchmark algorithmic classification tasks with transformers and LLMs. **The labelling progress is indicated in the coverage badge above.**

We are labelling two samples of novel datasets of political digital interactions on Twitter (rebranded as X). The first set comprises almost 5 million data points from three Latin American protest events: (a) protests against the coronavirus and judicial reform measures in Argentina during August 2020; (b) protests against education budget cuts in Brazil in May 2019; and (c) the social outburst in Chile stemming from protests against the underground fare hike in October 2019. We are focusing on interactions in Spanish to elaborate a gold standard for digital interactions in this language, therefore, we prioritise Argentinian and Chilean data. The second set contains more than 31 million messages and more than 9 million interactions between 2010 and 2022, covering the election of members of the first Constitutional Convention in Chile, the drafting process and the referendum in which the proposal was rejected.

This project is generously funded by the **OpenAI Academic Programme**, **2024 FAE-UDP Research Grant**, and partially by the **St Hilda's College Muriel Wise Fund at the University of Oxford**. The Training Data Lab research group also logistically supports this project.

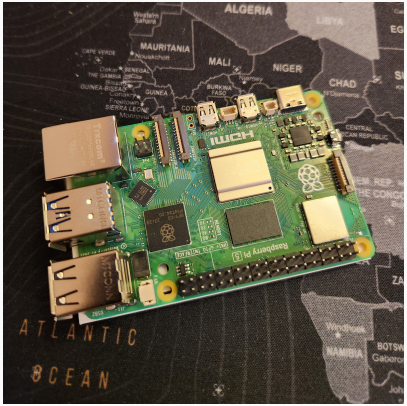
Brazil was excluded to compare, in the future, this benchmarking with the **gold standard** that is being created using Spanish native speakers' **human coders** in this project:

GitHub repository.

Raspberry Pi 5 Classification

We used Perspective API as a proxy of **ground-truth labels** and as a gold standard for our benchmarking. The classification implied **2,411 hours** of computing **(entire dataset, not only the sample)**.

The process was deployed on a **Raspberry Pi 5**, a small computer board with an ARM CPU and meagre power consumption. Our footprint using Raspberry Pi 5 was 6.51 kg of CO2, which implied a **reduction of 96% in emissions** that we would have generated running the tasks on an average cloud computing service.



* Picture of the Raspberry Pi 5 used for classification.

Open-Source LLMs for Text-Annotation

We selected our LLMs for benchmarking by running a **pre-proof-of-concept (pre-PoC)** using only a subsample of 100 random observations and **24 general-purpose models** deployed locally.

We discarded models that tend to provide an analysis of the comments rather than a straightforward annotation and those that took over one minute per observation.

Open-Source LLMs for Text-Annotation

Finally, we selected 13 LLMs and included some of their versions with more parameters; thus, we **benchmarked 16 open-source LLMs**. Our prompt strategy was based on the core definitions of Perspective for toxicity with the following system message:

System message

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation.

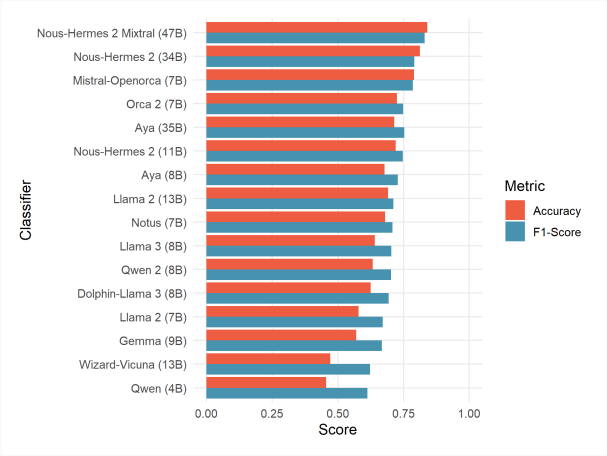
Along with providing texts of our balanced sample, we also listed the categories for the task as follows: **"Respond with only the category (TOXIC or NONTOXIC). Do not provide any additional analysis or explanation."**

Results

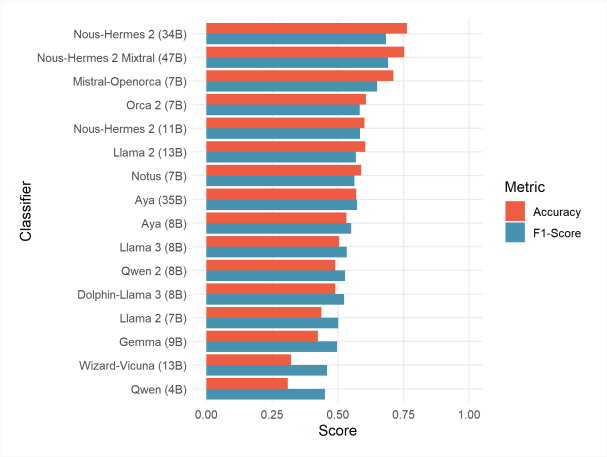
Error Rate Analysis (Toxicity at 0.55)

Nous-Hermes 2 with 34B parameters and its version trained on **Mixtral with 47B** parameters show the highest accuracy and F1-score in classifying toxicity for the more relaxed threshold at 0.55 and the standard at 0.70.

Nous-Hermes models were trained to excel at scientific discussions and coding tasks.



Error Rate Analysis (Toxicity at 0.70)

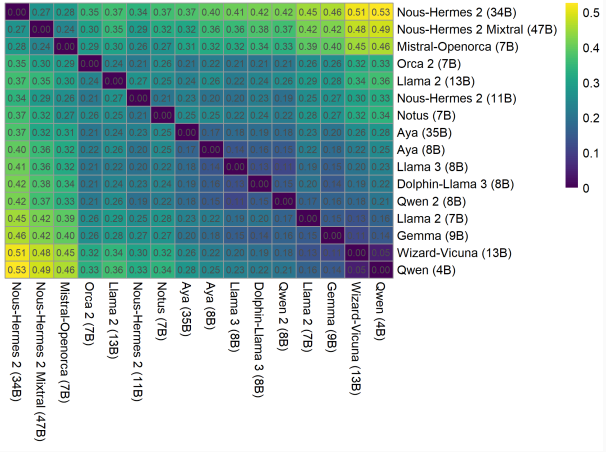


Mistral-Openorca and **Orca 2**, both with **7B parameters**, also show excellent performance that ranges between 60 and 79% depending on the models and the toxicity threshold for accuracy and slightly lower values for F1-score.

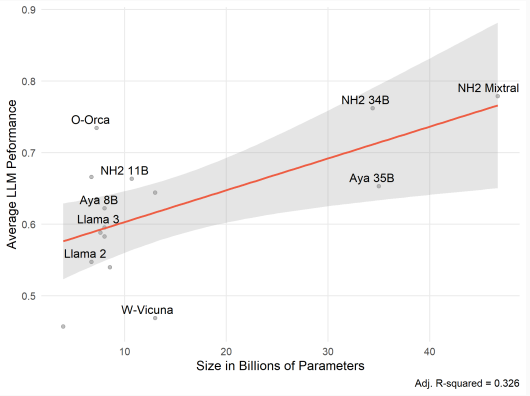
Jaccard Distance Zero-Shot LLMs Classifiers

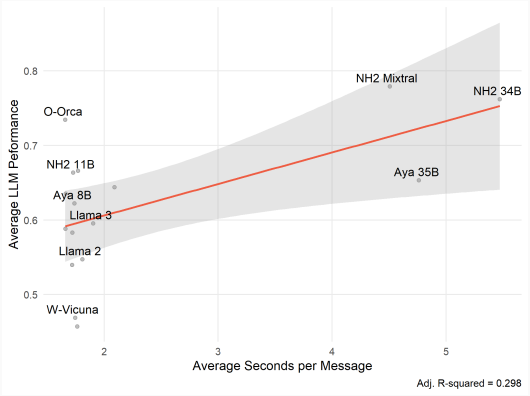
The trend is that the models with **worse performance tend to be more similar**, while interestingly, the models with the best performance show indexes around 0.30; therefore, they are not highly similar despite their common good performance.

This opens avenues for **ensemble LLMs annotations** and the use of stacking classifiers.



Performance, Parameters and Computing Time





Takeaways

Takeaways

☑ The models with better performance were those derived from **Nous-Hermes 2**, however, there is a **trade-off** between the number of parameters and computing time of the models and their accuracy.

☑ The only exception is **Mistral-Openorca**, which, with a small number of parameters, was able to classify toxicity with high levels of accuracy and faster than the largest models.

☑ **Limitations.** We need to incorporate proprietary models (e.g., OpenAI GPTs); our ground-truth labels are a proxy of a gold standard, notwithstanding that Perspective API offers a reliable baseline for comparison for toxicity.

☑ **New avenues.** Temperature experiments and hallucination measurement; more complex classifications (few-shot or chain-of-thought); and use this to analyse incivility dynamics on specific cases.

Acknowledgements and Funding

This work in progress is the first step in a broader project funded by the OpenAI Academic Programme and the Faculty of Administration and Economics at the Universidad Diego Portales, Chile. I also thank the Institute of Security and Global Affairs at Leiden University, Netherlands, and the Training Data Lab research group for their support.

udp UNIVERSIDAD DIEGO PORTALES

tDl

References

References I

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., and Gilardi, F. (2024). Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. Preprint, arXiv.

Bucher, M. J. J. and Martini, M. (2024). Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. Preprint, arXiv.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

González-Bustamante, B. (2023). *Critical events and ministerial turnover in Latin American presidential democracies*. PhD thesis, St Hilda's College, University of Oxford.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. (2024). AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. Preprint, arXiv.

Hopp, T., Vargo, C. J., Dixon, L., and Thain, N. (2020). Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept. *Social Science Computer Review*, 38(5):584–599.

References II

Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., Tang, L., Kang, Y., and Mars, J. (2024). Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. Preprint, arXiv.

Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6):922–946.

Orchard, X., González-Bustamante, B., and Cisternas, C. (2024). Amplification, contestation, and hostile engagement with news in social media: Comparing interactions with journalistic content before and after a protest cycle. In *Media and Governance in Latin America, Spain and Portugal*, Valencia, Spain. Universitat de València.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Schmidt, F., Stier, S., and Otto, L. (2024). Incivility in Comparison: How Context, Content, and Personal Characteristics Predict Exposure to Uncivil Content. *Social Science Computer Review*. OnlineFirst.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub